Brain Computer Interaction

Module III

Clustering techniques

Indian Institute of Information Technology
IIIT Sri City

Clustering techniques

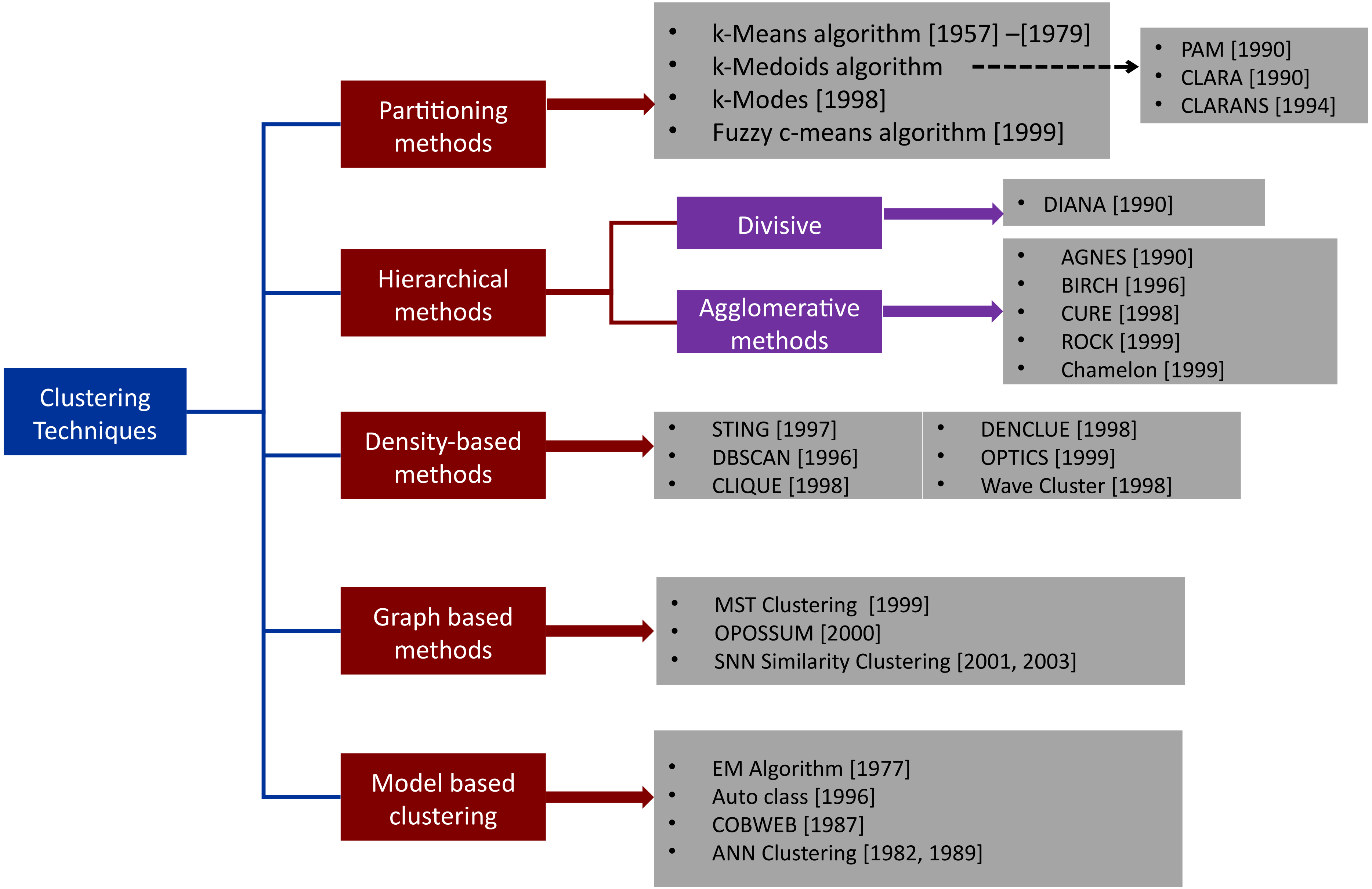
- Clustering has been studied extensively for more than 40 years and across many disciplines due to its broad applications.

- As a result, many clustering techniques have been reported in the literature.

- Let us categorize the clustering methods. In fact, it is difficult to provide a crisp categorization because many techniques overlap to each other in terms of clustering paradigms or features.

- A broad taxonomy of existing clustering methods is shown in the next slide.

- It is not possible to cover all the techniques in this lecture series. We emphasize on major techniques belong to partitioning and hierarchical algorithms.


Clustering Techniques
Partitioning methods
k-Means algorithm [1957] –[1979]
k-Medoids algorithm
k-Modes [1998]
Fuzzy c-means algorithm [1999]
PAM [1990]
CLARA [1990]
CLARANS [1994]
Hierarchical methods
Divisive
DIANA [1990]
Agglomerative methods
AGNES [1990]
BIRCH [1996]
CURE [1998]
ROCK [1999]
Chamelon [1999]
Density-based methods
STING [1997]
DBSCAN [1996]
CLIQUE [1998]
DENCLUE [1998]
OPTICS [1999]
Wave Cluster [1998]
Graph based methods
MST Clustering [1999]
OPOSSUM [2000]
SNN Similarity Clustering [2001, 2003]
Model based clustering
EM Algorithm [1977]
Auto class [1996]
COBWEB [1987]
ANN Clustering [1982, 1989]

k-Means Algorithm

- k-Means clustering algorithm proposed by J. Hartigan and M. A. Wong [1979].

- Given a set of n distinct objects, the k-Means clustering algorithm partitions the objects into k number of clusters such that intracluster similarity is high but the intercluster similarity is low.

- In this algorithm, user has to specify k, the number of clusters and consider the objects are defined with numeric attributes and thus using any one of the distance metric to demarcate the clusters.

k-Means Algorithm

The algorithm can be stated as follows.

- First it selects k number of objects at random from the set of n objects. These k objects are treated as the centroids or center of gravities of k clusters.

- For each of the remaining objects, it is assigned to one of the closest centroid. Thus, it forms a collection of objects assigned to each centroid and is called a cluster.

- Next, the centroid of each cluster is then updated (by calculating the mean values of attributes of each object).

- The assignment and update procedure is until it reaches some stopping criteria (such as, number of iteration, centroids remain unchanged or no assignment, etc.)

k-Means Algorithm

Algorithm 24.1: k-Means clustering

Input: D is a dataset containing *n* objects, *k* is the number of cluster

Output: A set of *k* clusters

Steps:

1. Randomly choose *k* objects from D as the initial cluster centroids.
2. **For** each of the objects in D **do**
   - Compute distance between the current objects and *k* cluster centroids
   - Assign the current object to that cluster to which it is closest.
3. Compute the "cluster centers" of each cluster. These become the new cluster centroids.
4. Repeat step 2-3 until the convergence criterion is satisfied
5. Stop

k-Means Algorithm

Note:

1) Objects are defined in terms of set of attributes. where each is continuous data type.
2) Distance computation: Any distance such as or cosine similarity.
3) Minimum distance is the measure of closeness between an object and centroid.
4) Mean Calculation: It is the mean value of each attribute values of all objects.
5) Convergence criteria: Any one of the following are termination condition of the algorithm.
    - Number of maximum iteration permissible.
    - No change of centroid values in any cluster.
    - Zero (or no significant) movement(s) of object from one cluster to another.
    - Cluster quality reaches to a certain level of acceptance.

Illustration of k-Means clustering algorithms

Table 24.1: 16 objects with two attributes and .

A_1	A_2
6.8	12.6
0.8	9.8
1.2	11.6
2.8	9.6
3.8	9.9
4.4	6.5
4.8	1.1
6.0	19.9
6.2	18.5
7.6	17.4
7.8	12.2
6.6	7.7
8.2	4.5
8.4	6.9
9.0	3.4
9.6	11.1

Fig 24.1: Plotting data of Table 24.1

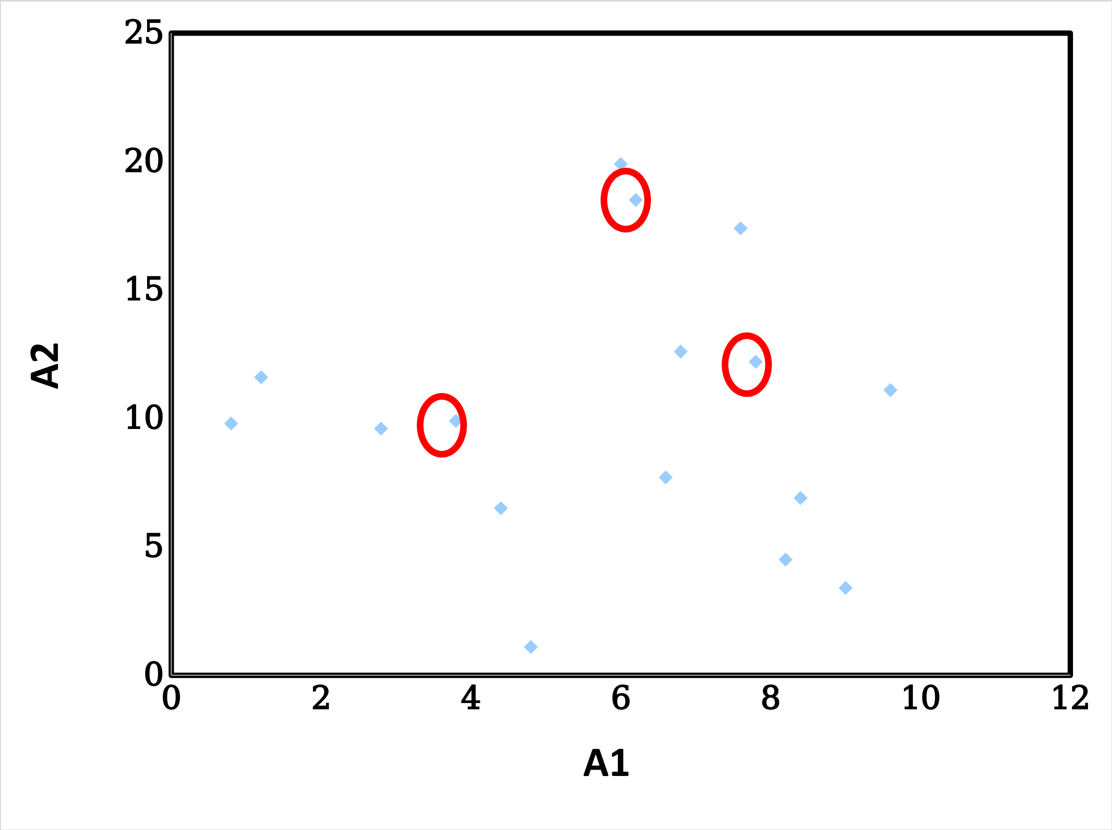


Illustration of k-Means clustering algorithms

- Suppose, k=3. Three objects are chosen at random shown as circled (see Fig 24.1). These three centroids are shown below.

Initial Centroids chosen randomly

Centroid	Objects	
c_1	3.8	9.9
c_2	7.8	12.2
c_3	6.2	18.5

- Let us consider the Euclidean distance measure (L_2 Norm) as the distance measurement in our illustration.
- Let d_1, d_2 and d_3 denote the distance from an object to c_1, c_2 and c_3 respectively. The distance calculations are shown in Table 24.2.
- Assignment of each object to the respective centroid is shown in the right-most column and the clustering so obtained is shown in Fig 24.2.

Illustration of k-Means clustering algorithms

Table 24.2: Distance calculation

A_1	A_2	d_1	d_2	d_3	cluster
6.8	12.6	4.0	1.1	5.9	2
0.8	9.8	3.0	7.4	10.2	1
1.2	11.6	3.1	6.6	8.5	1
2.8	9.6	1.0	5.6	9.5	1
3.8	9.9	0.0	4.6	8.9	1
4.4	6.5	3.5	6.6	12.1	1
4.8	1.1	8.9	11.5	17.5	1
6.0	19.9	10.2	7.9	1.4	3
6.2	18.5	8.9	6.5	0.0	3
7.6	17.4	8.4	5.2	1.8	3
7.8	12.2	4.6	0.0	6.5	2
6.6	7.7	3.6	4.7	10.8	1
8.2	4.5	7.0	7.7	14.1	1
8.4	6.9	5.5	5.3	11.8	2
9.0	3.4	8.3	8.9	15.4	1
9.6	11.1	5.9	2.1	8.1	2

Fig 24.2: Initial cluster with respect to Table 24.2

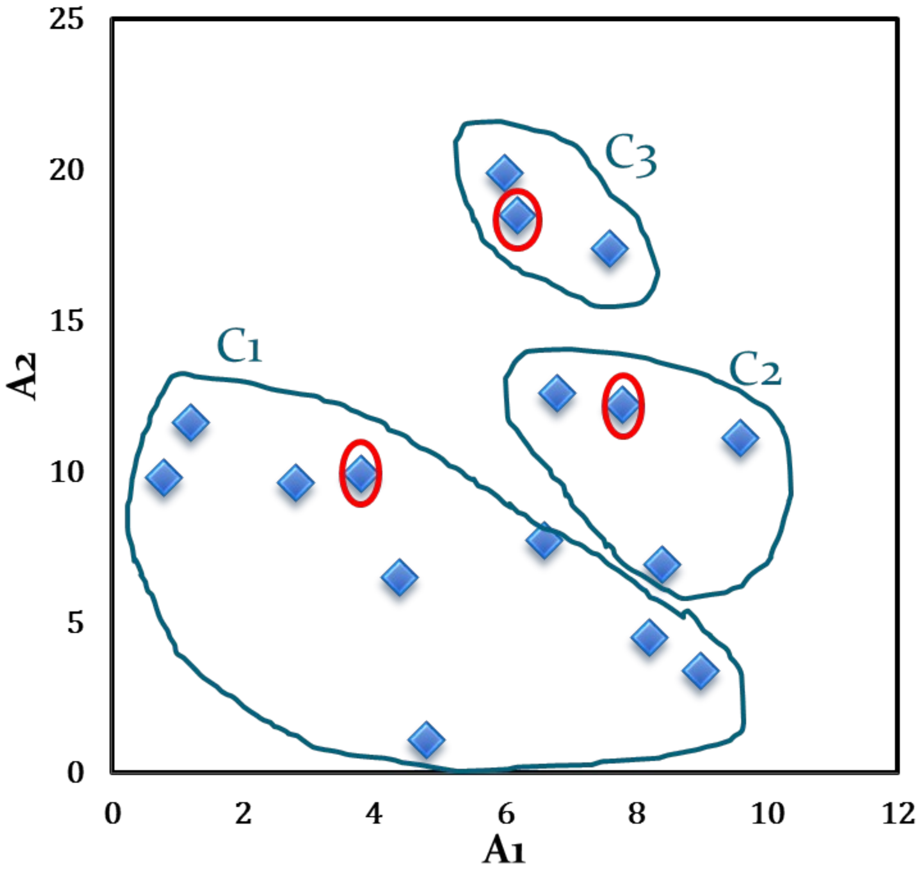


Illustration of k-Means clustering algorithms

The calculation new centroids of the three cluster using the mean of attribute values of A_1 and A_2 is shown in the Table below. The cluster with new centroids are shown in Fig 24.3.

Calculation of new centroids

New Centroid	Objects	
c_1	4.6	7.1
c_2	8.2	10.7
c_3	6.6	18.6

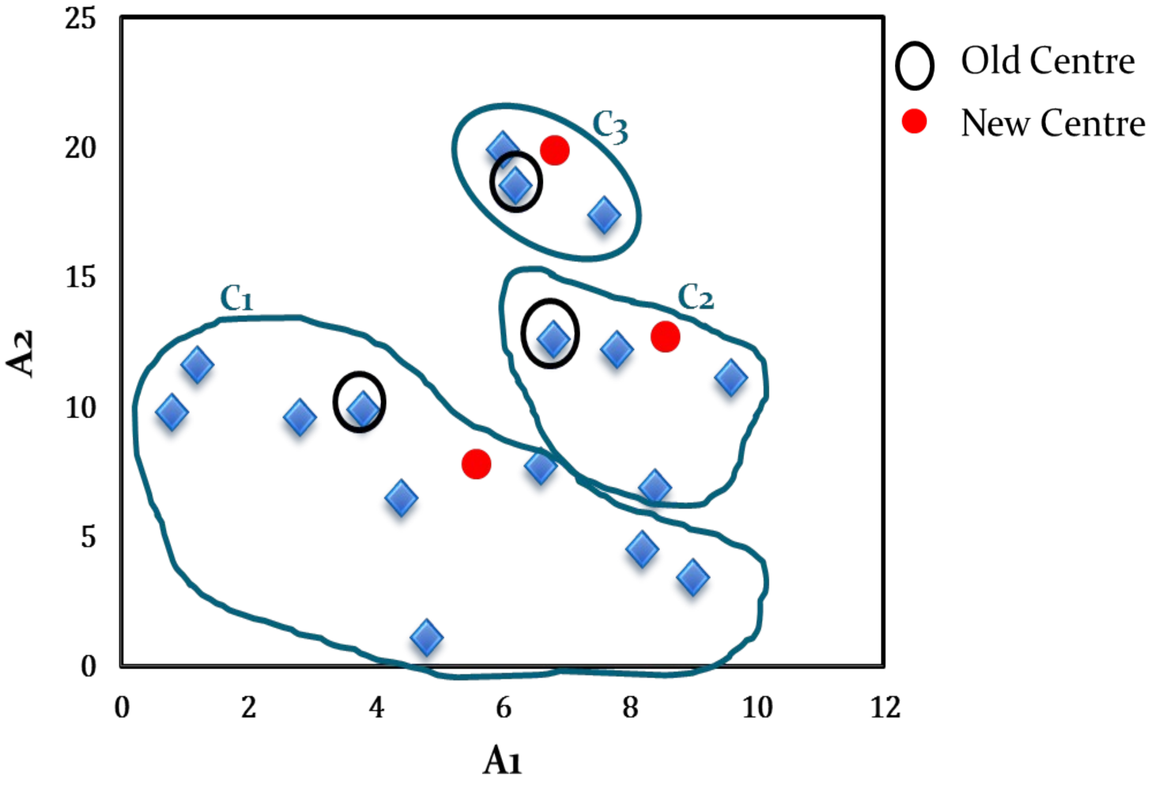


Fig 24.3: Initial cluster with new centroids

Illustration of k-Means clustering algorithms

We next reassign the 16 objects to three clusters by determining which centroid is closest to each one. This gives the revised set of clusters shown in Fig 24.4.

Note that point p moves from cluster C_2 to cluster C_1.

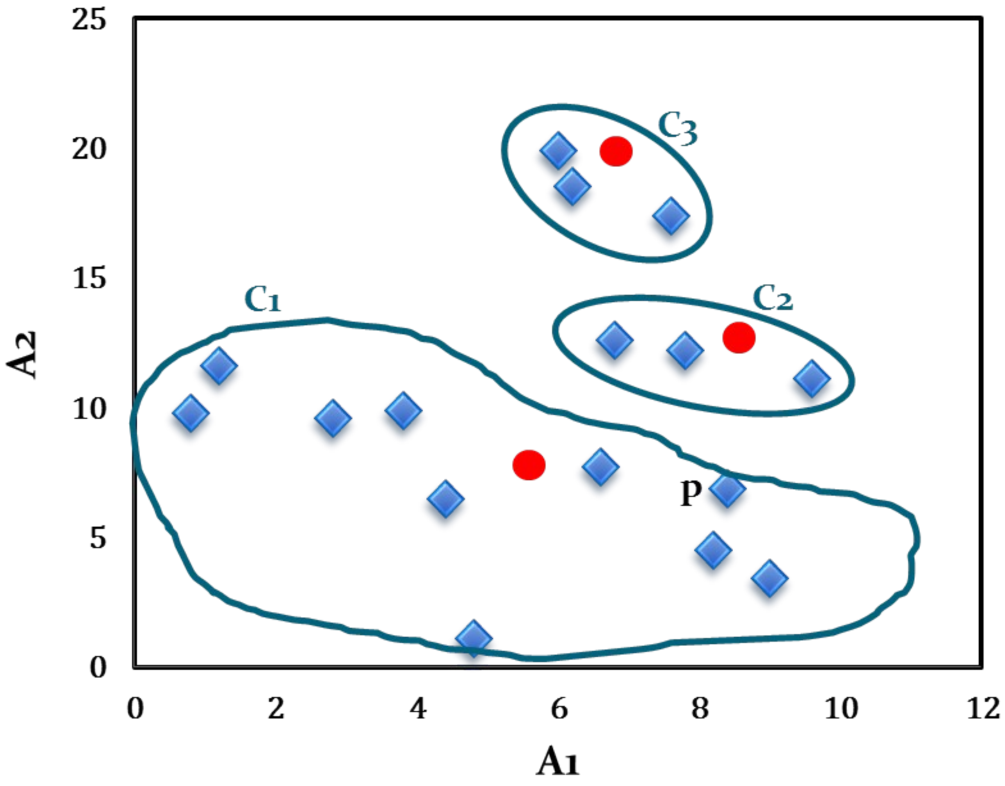


Fig 24.4: Cluster after first iteration

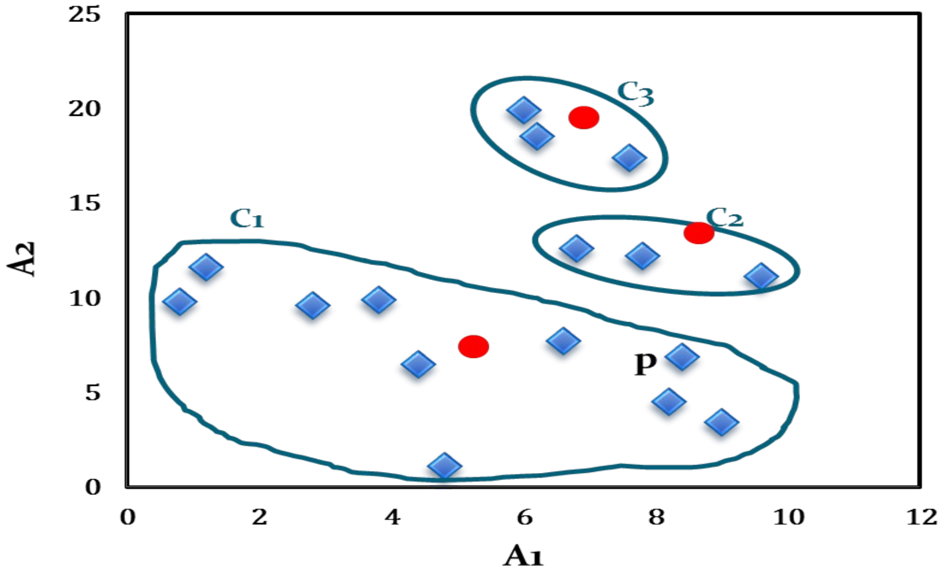
Illustration of k-Means clustering algorithms

- The newly obtained centroids after second iteration are given in the table below. Note that the centroid c_3 remains unchanged, where c_2 and c_1 changed a little.
- With respect to newly obtained cluster centres, 16 points are reassigned again. These are the same clusters as before. Hence, their centroids also remain unchanged.
- Considering this as the termination criteria, the k-means algorithm stops here. Hence, the final cluster in Fig 24.5 is same as Fig 24.4.

Cluster centres after second iteration

Centroid	Revised Centroids	
c_1	5.0	7.1
c_2	8.1	12.0
c_3	6.6	18.6

Fig 24.5: Cluster after Second iteration



Comments on k-Means algorithm

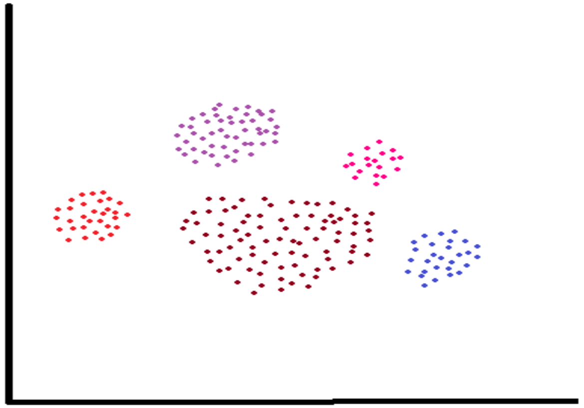
Advantages:

- k-Means is simple and can be used for a wide variety of object types.
- It is also efficient both from storage requirement and execution time point of views. By saving distance information from one iteration to the next, the actual number of distance calculations, that must be made can be reduced (specially, as it reaches towards the termination).

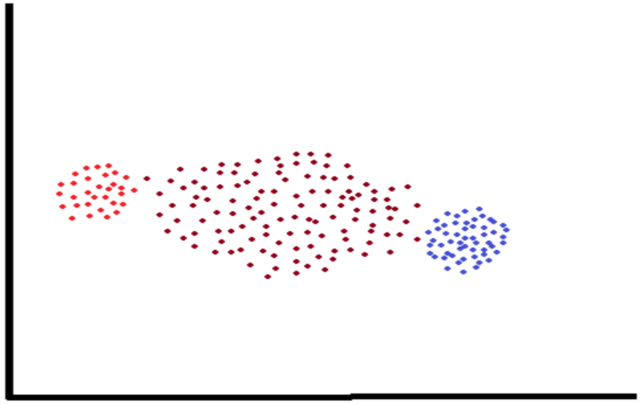
Limitations:

- The k-Means is not suitable for all types of data. For example, k-Means does not work on categorical data because mean cannot be defined.
- k-means finds a local optima and may actually minimize the global optimum.
- k-means has trouble clustering data that contains outliers.
- k-Means algorithm cannot handle non-globular clusters, clusters of different sizes and densities (see Fig 24.6 in the next slide).
- k-Means algorithm not really beyond the scalability issue (and not so practical for large databases).

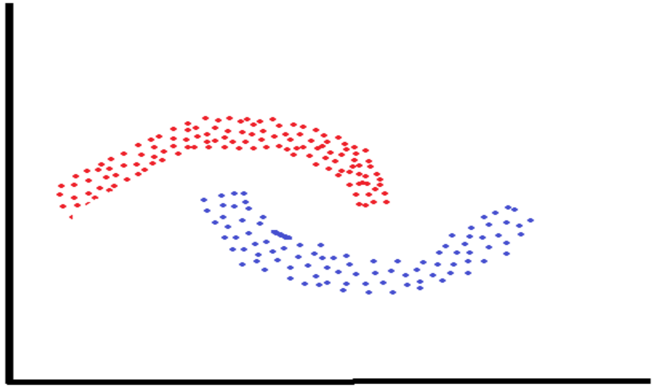
Comments on k-Means algorithm



Cluster with different sizes



Cluster with different densities



Non-convex shaped clusters

Fig 24.6: Some failure instance of k-Means algorithm

Different variants of k-means algorithm

There are a quite few variants of the k-Means algorithm. These can differ in the procedure of selecting the initial *k* means, the calculation of proximity and strategy for calculating cluster means. Another variants of k-means to cluster categorical data.

Few variant of k-Means algorithm includes

- Bisecting k-Means (addressing the issue of initial choice of cluster means).
    1. M. Steinbach, G. Karypis and V. Kumar "A comparison of document clustering techniques", *Proceedings of KDD workshop on Text mining,* 2000.
- Mean of clusters (Proposing various strategies to define means and variants of means).
    - B. zhan "Generalised k-Harmonic means – Dynamic weighting of data in unsupervised learning", *Technical report, HP Labs*, 2000.
    - A. D. Chaturvedi, P. E. Green, J. D. Carroll, "k-Modes clustering", *Journal of classification*, Vol. 18, PP. 35-36, 2001.
    - D. Pelleg, A. Moore, "x-Means: Extending k-Means with efficient estimation of the number of clusters", *17^{th} International conference on Machine Learning,* 2000.

Different variants of k-means algorithm

- N. B. Karayiannis, M. M. Randolph, “Non-Euclidean c-Means clustering algorithm”, *Intelligent data analysis journal*, Vol 7(5), PP 405-425, 2003.

- V. J. Olivera, W. Pedrycy, “Advances in Fuzzy clustering and its applications”, Edited book. John Wiley [2007]. (Fuzzy c-Means algorithm).

- A. K. Jain and R. C. Bubes, “Algorithms for clustering Data”, Prentice Hall, 1988.

  Online book at http://www.cse.msu.edu/~jain/clustering_Jain_Dubes.pdf

- A. K. Jain, M. N. Munty and P. J. Flynn, “Data clustering: A Review”, *ACM computing surveys*, 31(3), 264-323 [1999]. Also available online.

Any question?